Detection of Pediatric Respiratory and Gastrointestinal Outbreaks from Free-Text Chief Complaints

Oleg Ivanov M.D. M.P.H., Per H. Gesteland M.D., William Hogan M.D., Michael B. Mundorff M.B.A. / H.S.A. and Michael M. Wagner M.D. Ph.D.

The RODS Laboratory, Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
University of Utah and Intermountain HealthCare, Salt Lake City, UT

We conducted a retrospective study to ascertain the potential of free-text chief complaints collected in pediatric emergency departments to serve as surveillance data for early detection of outbreaks.

We determined that automatically coded chief complaint data provide a signal that reflects outbreaks in a population of children less than five years of age. Using the Exponentially Weighted Moving Average (EWMA) detection algorithm, we measured the timeliness, sensitivity, and specificity of free-text chief complaints for predicting outbreaks of pediatric respiratory and gastrointestinal illness.

We found that time series of automatically coded free text-chief complaints in pediatric patients correlate well with hospital admissions and precede them by the mean of 10.3 days (95% CI -15.15, 35.5) for respiratory outbreaks and 29 days (95% CI 4.23, 53.7) for gastrointestinal outbreaks.

We conclude that free-text chief complaints may play an important role as an early, sensitive and specific indicator of outbreaks of respiratory and gastrointestinal illness in children less than five years of age.

INTRODUCTION

Timeliness is a crucial characteristic of an early-warning biosurveillance system. Early detection of an outbreak leads to its early containment and that translates into saved lives. One way to improve timeliness of outbreak detection is to use alternative data sources as suggested by Wagner et al. [1] Among the proposed alternative data sources for early outbreak detection are school or industrial absenteeism, retail over-the-counter and pharmacy purchases, web queries, and data collected during emergency room visits. Such sources of data form the basis of new biosurveillance systems being developed in several locations. [2,3]

However, relatively little work has been done to measure the performance characteristics of biosurveillance systems that are based on such data. Only two prior studies have measured the sensitivity, specificity, and timeliness of influenza outbreak detection from such data and both suffered from either methodological limitations or very small sample size. [4,5]

The present study attempts to measure performance characteristics of a biosurveillance system based on free-text chief complaints, a type of data collected during emergency room visits. Currently the scientific evidence as to the value of free-text chief complaints for outbreak detection is limited to the results of one prior study. [6] In that study, we measured case detection accuracy of free-text chief complaint classifiers and demonstrated that a case of acute infectious gastrointestinal illness can be detected by automatic coding of a chief complaint with sensitivity 0.63 and specificity 0.94. The fact that the focus of the study was limited to case detection accuracy did not allow us to make any conclusions regarding the value of free-text chief complaints for outbreak detection. To our knowledge there are no studies that measured sensitivity, specificity and timeliness of outbreak detection from automatically coded free-text chief complaint data relative to other types of surveillance data.

The primary hypothesis in the present study is that automatically coded chief complaint data provide a signal that reflects an outbreak in a certain population.

The secondary hypothesis is that chief complaint data are an early indicator of outbreaks relative to inpatient admission data.

METHOD

We studied pediatric free-text chief complaints and hospital admissions for patients residing in four contiguous counties (Weber, Davis, Salt Lake, Utah) in Utah collectively known as the Wasatch Front area. Over 80% of the state’s population resides in these counties.

Free-text Chief Complaints

A chief complaint is a free-text string representing a mixture of subjective and objective information describing a patient’s status on his/her visit to an emergency department. It is recorded by an

Figure 1. Respiratory time series

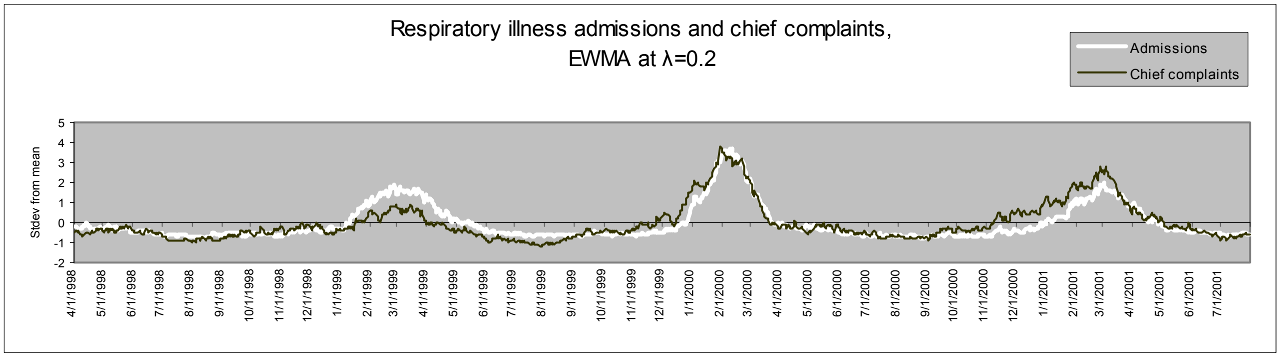
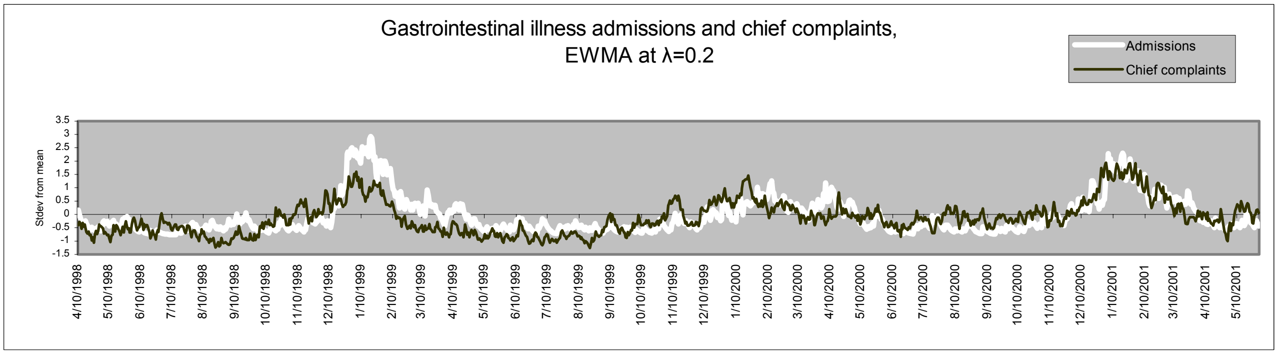


Figure 2. Gastrointestinal time series



admission clerk during patient registration. We consider chief complaints to be one of the earliest elements of clinical information available to automated systems.

Primary Children's Medical Center (PCMC), Salt Lake City, Utah provided 4 years (1998-2001) of free-text chief complaints. PCMC is a tertiary referral hospital for the study region. The chief complaints were classified with a Naïve Bayes classifier implemented in the publicly available Complaint Coder (CoCo) software. [7] A training dataset for the classifier was created by physician review and categorization of free-text chief complaints. The classifier categorizes a free-text chief complaint into one of 8 categories or syndromes. For the purpose of this study, we restricted our case definition to patients under age of five years whose chief complaints were classified into respiratory or gastrointestinal syndrome. We aggregated the respiratory and gastrointestinal syndrome cases into daily counts based on their admission date. Then we rescaled the obtained time series by computing number of standard deviations of an observation from the mean.

Gold Standard Outbreaks

We obtained a dataset derived from the Utah Hospital Discharge Database for the years 1998-2001 inclusive. The Utah Hospital Discharge Database includes all hospital discharge diagnoses for all hospitals in the state coded using ICD-9 CM.

We created sets of ICD-9 codes that we derived from the CDC pneumonia and influenza mortality surveillance system set and from knowledge of pediatric infectious disease. We considered two groups of illnesses that present as seasonal outbreaks in the study population – infectious lower respiratory tract illness due to pneumonia, influenza and bronchiolitis; and infectious gastrointestinal illness due to rotavirus or other causes of pediatric gastroenteritis. Therefore our code sets included ICD-9 codes used to encode discharge diagnoses of the following four diseases of interest: pneumonia and influenza, bronchiolitis, rotavirus, and pediatric gastroenteritis.

Using these code sets we created time series of daily counts of admissions for patients under age of five years with the corresponding discharge diagnoses. In creating the time series, we used date of admission rather than date of discharge because we wanted the experiment to be a rigorous test of the timeliness of chief complaints data, and the use of discharge date would have biased the timeliness measure by inclusion of patients' hospital length of stay. The obtained time series were rescaled using approach described above for chief complaint data.

Table 1. Timeliness measurements with EWMA detection algorithm and sensitivity analysis

Syndrome	Outbreak period	Threshold* multiplier, k	Max timeliness, days	Min timeliness, days	Mean timeliness, days
Respiratory	1998-99	6	-7	-18	-11.7
Respiratory	1999-00	13	16	7	10
Respiratory	2000-01	12	50	26	32.7
GI	1998-99	5	38	13	34.8
GI	1999-00	8	80	27	47.5
GI	2000-01	3	24	-5	4.9

* The first threshold that did not produce false positives

Measurement of Timeliness

We measured timeliness in two ways. First, we computed the cross-correlation of the two time series and looked at the time lag that maximized that function. [8]

The second method for measuring timeliness utilized a detection algorithm. We ran the algorithm on both the gold standard discharge diagnosis time series as well as the chief complaint data. Timeliness was determined as the difference between the date the algorithm first fired an alert on the gold standard and the date the algorithm first fired on the chief complaint data at the lowest alert threshold that did not yield false positive alerts. A false positive alert was defined as an alert in non-outbreak period. Non-outbreak periods were determined by expert review of time series. A non-outbreak period was defined as a period of low in-patient admission counts with low variability. The specific detection algorithm we used was the Exponentially Weighted Moving Average (EWMA) implemented in SAS/QC software package. EWMA is a smoothing technique, which calculates a weighted value of a current observation with the following formula:

$$E_t = \lambda * X_t + (1 - \lambda) * E_{t-1},$$

where E_t is a current weighted ("smoothed") value, X_t is a current observed value, E_{t-1} is a weighted ("smoothed") value at time t-1 and λ is a weighting ("smoothing") parameter determining the degree to which 'older' data enter into the calculation of the smoothed value. A larger value of λ gives more weight to recent data and less weight to older data; a smaller value of λ gives more weight to older data. The value of λ is usually set between 0.2 and 0.3 but this choice is considered arbitrary. Note that at $\lambda = 1$ no smoothing of the data occurs. [9,10]

The detection algorithm also requires setting an upper control limit that is calculated from the process mean and standard deviation of a time series:

$$UCL= \mu+k*\sigma*\mathrm{sqrt}(\lambda/n*(2 - \lambda)),$$

where UCL is an upper control limit, μ is a time series mean, *k* is an integer multiplier, σ is a time series or process standard deviation, and n is the length of a time series. An alarm is produced when the "smoothed" signal crosses the upper control limit. Parameter *k* is arbitrarily chosen based on the desired control limit height and therefore controls the tradeoff between sensitivity and specificity of the algorithm. Under the assumption that the monitoring starts in a non-outbreak period, we computed the mean and the standard deviation of a time series in non-outbreak periods. Since parameters λ and k are specified arbitrarily we decided to test whether the measurements of timeliness, sensitivity and specificity were sensitive to the particular selection of parameter λ and *k*. We varied the value of λ between 0.3 and 0.01. The upper bound of 0.3 provides a minimum recommended degree of smoothing required for reliable performance of the detection algorithm. The lower bound of 0.01 provides a level of smoothing at which the key features of the signals are not yet "smoothed" out. We limited the analysis of timeliness to values of *k* that produced no false alarms because false alarms cause erroneous measurements of timeliness. We measured specificity during non-outbreak periods. We defined false positive alarms as those that occurred during non-outbreak periods.

RESULTS

Respiratory illness and syndrome

The average number of admissions per day that resulted in the assignment of a discharge code from the pneumonia, influenza and bronchiolitis set was six with a maximum of 43 and a minimum of zero.

The average number of visits per day to PCMC emergency department with chief complaints classified by CoCo as respiratory syndrome cases was 10.3 with a maximum of 61 and a minimum of zero.

The mean timeliness calculated across different values of λ varied from –11.7 to 32.7 days with an

Table 2. Sensitivity and specificity at different threshold levels, EWMA at $\lambda = 0.2$

Respiratory outbreaks			Gastrointestinal outbreaks		
Threshold, k	Sensitivity	Specificity	Threshold, k	Sensitivity	Specificity
1 – 4	1.0 (3/3)	0.0 (0/3)	1 – 2	1.0 (3/3)	0.0 (0/3)
5 – 8	1.0 (3/3)	0.33 (1/3)	3	1.0 (3/3)	0.33 (1/3)
9	1.0 (3/3)	0.66 (2/3)	4	1.0 (3/3)	0.66 (2/3)
10 – 12	1.0 (3/3)	1.0 (3/3)	5 – 9	1.0 (3/3)	1.0 (3/3)
13 – 18	0.66 (2/3)	1.0 (3/3)	10	0.66 (2/3)	1.0 (3/3)
19	0.33 (1/3)	1.0 (3/3)	11	0 (0/3)	1.0 (3/3)

overall mean of 10.3 days (95% CI -15.15, 35.5) (see Table 1).

Sensitivity and specificity of outbreak detection from chief complaints reached 1.0 at *k* = 10 through 12 (see Table 2).

Gastrointestinal illness and syndrome

The average number of admissions per day that resulted in the assignment of a discharge code from the gastrointestinal illness set was one with a maximum of ten and a minimum of zero.

The average number of visits per day to PCMC emergency department with chief complaints classified by CoCo as gastrointestinal syndrome cases was 6.3 with a maximum of 22 and a minimum of zero.

The mean timeliness calculated across different values of λ varied from 4.9 to 47.5 days with an overall mean of 29 days (95% CI 4.23, 53.7) (see Table 1).

Sensitivity and specificity of outbreak detection from chief complaints reached 1.0 at *k* = 5 through 9 (see Table 2).

Correlation Coefficient

Cross-correlation analysis of three respiratory outbreaks resulted in a mean timeliness of 7.4 days (95% C.I. –8.34, 43.3). In the case of three gastrointestinal outbreaks the mean timeliness was 17.6 days (95% C.I. 3.4, 46.7).

DISCUSSION

We consider free-text chief complaints to be a very important biosurveillance data source due to their ubiquity and real-time availability in electronic format. The results of the present study demonstrate that syndrome counts obtained by automatic classification of free-text chief complaints correlate very closely with disease activity as measured by hospital admissions for both gastrointestinal and respiratory illness in children. This provides evidence in support of the primary hypothesis. The presence of a set of detection thresholds that allowed for the detection of both types of outbreaks from chief complaint data with 100% sensitivity and specificity (see Table 2) suggests that the strength of the signal in chief complaint time series allows them to be used as a data source in a sensitive and specific biosurveillance system, under conditions of less than perfectly accurate probabilistic case detection. However, in this study, we observed relatively large outbreaks and this statement might not hold for outbreaks of smaller size. Determination of the smallest outbreak detectable from probabilistically coded chief complaints is future work.

The mean timeliness of respiratory outbreak detection from chief complaints was 10.3 days (95% CI -15.15, 35.5). The result of the cross-correlation analysis was consistent with this result. The 95% confidence intervals include values less than zero. This result is due to the 1998-99 outbreak period in which the increase in admissions for respiratory illness preceded the increase in number of respiratory syndrome cases (see Figure 1). Since we defined the cases of respiratory illness based on their discharge diagnosis and aggregated the counts of cases based on their admission date, one explanation of this phenomenon may be the presence of a nosocomial outbreak, which could have resulted in artificially high counts of admissions due to respiratory illness while in reality majority of the patients could have been admitted for a different reason, when they had no respiratory infection. It is known that RSV - the major cause of bronchiolitis - often causes pediatric nosocomial infections. [11,12] Another possible explanation is interaction between respiratory and gastrointestinal outbreaks in a single community. Figure 2 shows a pronounced gastrointestinal illness outbreak immediately preceding the respiratory outbreak in the 1998-1999 season. Such an outbreak could result in saturation of capacity at PCMC and affect community referral patterns. An increase in referrals to facilities other than PCMC would attenuate the magnitude of change in visit counts with respiratory chief complaints (that we were able to obtain for PCMC only) and decrease the measurement of timeliness.

The mean timeliness of gastrointestinal outbreak detection from chief complaints was 29 days (95% CI 4.23, 53.7). The result of the cross-correlation analysis was consistent with this result. The 95% confidence intervals do not include zero therefore we conclude that the increase in visits with gastrointestinal chief complaints consistently precedes the increase in admissions for gastrointestinal illness.

There are several unique contributions that this study makes to the field of biosurveillance and, more specifically, to the field of automated early outbreak detection. First, it establishes a framework for determining the value of different types of alternative data sources for outbreak detection. Second, it demonstrates a potential value of automatically coded free-text chief complaints for early real-time outbreak detection. Third, it introduces a novel method of timeliness measurement of an alternative data source relative to an established biosurveillance data source such as discharge diagnosis data. The cross-correlation analysis—the traditional approach to measuring time difference and correlation between signals—has several disadvantages compared to the detection algorithm based approach described in this paper. Among these disadvantages are sensitivity to wide variations in the amplitude of time series and, more important, inability to provide any information about sensitivity and specificity of outbreak detection.

The limitations of this study included the following. We analyzed the value of free-text chief complaints for pediatric outbreak detection only, therefore the results and the conclusions may not generalize to the adult population. The types of outbreaks that we considered were relatively large seasonal outbreaks of respiratory and gastrointestinal illness. We had no ability to analyze the value of chief complaints for detection of outbreaks of smaller size and those due to other types of illnesses. We believe that such analyses should be conducted for the purpose of having better understanding of the value of alternative data sources for outbreak detection. The sample of only three outbreak periods for each illness as well as limitations described above do not allow us to make definitive conclusions as to the value of chief complaints for outbreak detection. We are currently conducting a multicenter study that is designed to address these limitations.

CONCLUSIONS

We conclude that automatically coded free-text chief complaints may play an important role as an early, sensitive and specific indicator of outbreaks of respiratory and gastrointestinal illness in children less than five years of age.

ACKNOWLEDGMENTS

This work was supported by grant T15 LM/DE07059 from the National Library of Medicine; contract 290-00-0009 from the Agency for Healthcare Research and Quality; and contract F30602-01-2-0550 sponsored by the Defense Advanced Research Projects Agency and managed by Rome Laboratory.

REFERENCES

1. Wagner M, Tsui F-C, Espino J, Dato V, Sittig D, Caruana R, et al. The emerging science of very early detection of disease outbreaks. J Public Health Manag Pract 2001;6(6):50-58.
2. Lober WB, Thomas Karras B, Wagner MM, Marc Overhage J, Davidson AJ, Fraser H, Trigg LJ, Mandl KD, Espino JU, Tsui FC. Roundtable on Bioterrorism Detection: Information System-based Surveillance. J Am Med Inform Assoc. 2002 Mar-Apr;9(2):105-15.
3. http://www.nytimes.com/2003/01/27/national/27DISE.html?ex=1044690552
4. Quenel P, Dab W, Hannoun C, Cohen JM. Sensitivity, specificity and predictive values of health service based indicators for the surveillance of influenza A outbreaks. Int J Epidemiol 1994;23(4):849-55.
5. Tsui F-C, Wagner MM, Dato V, Chang C-CH. Value of ICD-9-coded chief complaints for detection of outbreaks. Proc AMIA Symp 2001:711-715.
6. Ivanov O. et al. Accuracy of Three Classifiers of Acute Gastrointestinal Syndrome for Syndromic Surveillance, Proc AMIA Symp 2002: 345-349.
7. Olszewski R. T. Bayesian Classification of Triage Diagnoses for the Early Detection of Outbreaks. To appear in Proceedings of FLAIRS 2003.
8. Yaffee R. A. Introduction to Time Series Analysis and Forecasting. 2000 Academic Press, pp. 370–380
9. Hunter J. S. The Exponentially Weighted Moving Average. 1986 Journal of Quality Technology (18), pp. 203–210.
10. http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc324.htm
11. Gendrel D. et al. Coincidental outbreaks of rotavirus and respiratory syncytial virus in Paris: a survey from 1993 to 1998. Arch Pediatr 1999 Jul;6(7):735-9.
12. Goldmann DA. Epidemiology and prevention of pediatric viral respiratory infections in health-care institutions. Emerg Infect Dis 2001 Mar-Apr;7(2):249-53.